Neural Text Rewriting

Lai, Huiyuan

DOI:
10.33612/diss.910519765

CHAPTER 11

Responsibility Perspective Transfer

Different ways of linguistically expressing the same real-world event can lead to different perceptions of what happened. Previous work has shown that different descriptions of gender-based violence (GBV) influence the reader's perception of who is to blame for the violence, possibly reinforcing stereotypes which see the victim as partly responsible, too. As a contribution to raise awareness on perspective-based writing, and to facilitate access to alternative perspectives, we introduce the novel task of automatically rewriting GBV descriptions as a means to alter the perceived level of blame on the perpetrator. We present a quasi-parallel dataset of sentences with low and high perceived responsibility levels for the perpetrator, and experiment with unsupervised (mBART-based), zero-shot and few-shot (GPT3-based) methods for rewriting sentences. We evaluate our models using a questionnaire study and a suite of automatic metrics.

Chapter adapted from:

Minnema, G.*, Lai, H.*, Muscato, B., and Nissim, M. (2023). Responsibility perspective transfer for Italian femicide news. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7907–7918, Toronto, Canada. Association for Computational Linguistics

11.1 Introduction

"*A terrible incident involving husband and wife*", "*Husband kills wife*", "*Her love for him became fatal*": these different phrasings can all be used to describe the same violent event, in this case a *femicide*, but they won't trigger the same perceptions in the reader. Perceptions vary from person to person, of course, but also depend substantially and systematically on the different ways the same event is framed (Iyengar, 1994). Especially in the context of gender-based violence (GBV), this has important consequences on how readers will attribute responsibility: victims of femicides are often depicted, and thus perceived, as (co-)responsible for the violence they suffer.[1]

There is indeed evidence from the linguistic literature (Pinelli and Zanchi, 2021; Meluzzi et al., 2021) that people perceive responsibility differently according to how femicides are reported (more blame on the perpetrator in "*Husband kills wife*", more focus on the victim in "*Her love for him became fatal*"). In general, linguistic strategies that background perpetrators have been shown to favour victim blaming (Huttenlocher et al., 1968; Henley et al., 1995; Bohner, 2002; Gray and Wegner, 2009; Hart and Fuoli, 2020; Zhou et al., 2021b). This way of reporting contributes to reinforcing such social stereotypes.

If we want social stereotypes to be challenged, the language we use to describe GBV is thus an excellent place to start, also from a Natural Language Processing (NLP) perspective. Recent work has shown that perspectives on femicides and their triggered perceptions can be automatically modelled, e.g. perspectives discovery modelling (Minnema et al., 2022b) and perceived responsibility prediction (Minnema et al., 2022a). In this chapter, with the definition and examples shown in Table 11.1, we explore the challenge of *rewriting* descriptions of GBV with the aim to increase the perceived level of blame on the perpetrator, casting it as a text style transfer task (more details can be found in Chapter 3). In this novel *responsibility perspective transfer* task,

[1] A report on femicides from November 2018 by two Italian research institutes points out that the stereotype of a shared responsibility between the victim and its perpetrator is still widespread among young generations: "56.8% of boys and 38.8% of girls believe that the female is at least partly responsible for the violence she has suffered" (Laboratorio Adolescenza and Istituto IARD, 2018).

Definition	Given a sentence S that references an act of violence, write a sentence S' that describes the same facts as S but increases the perceived level of responsibility on the perpetrator of the violence.
Examples	*"A fatal stabbing"* $\rightarrow$ *"Someone stabbed another person to death"* *"Woman murdered by husband"* $\rightarrow$ *"Man murders wife"*
	metadata: name of the victim = Loredana Colucci, name of the perpetrator = Mohamed Aziz El Mountassir, relationship = ex-spouse, murder weapon = knife, municipality = Albenga

Table 11.1: Task definition.

a given sentence from femicide news reports gets rewritten in a way that puts more responsibility on the perpetrator, while preserving the original content.

Contributions We create an evaluation set containing quasi-aligned sentence pairs with "low" and "high" perceived levels of blame on perpetrators, expressing similar information relative to an event, from an existing dataset of Italian news. In absence of parallel training data, we train an unsupervised style transfer model introduced in Chapter 5 on non-parallel data (with style labels), as well as a zero-shot and a few-shot approach using the large language model GPT-3 (Brown et al., 2020) to perform rewriting. We run both human-based and automatic evaluations to assess the impact of rewriting on the perceived blame, comparing original and rephrased texts and find that models can achieve detectable perspective shifts. By introducing the novel task of responsibility perspective transfer, providing an evaluation dataset, a battery of trained models, and evidence of a successful methodology, we hope to foster further research and application developments on this and other perspective rewriting tasks that are relevant to society.

11.2 Datasets

In order to create an evaluation dataset for the task of responsibility perspective transfer, we make use of the *RAI femicide corpus* (Belluati, 2021), a dataset of 2,734 news articles covering 582 confirmed femicide cases and

198 other GBV-related cases[2] in Italy between 2012-2017. Of these, 182 cases (comprising 178 femicides and 4 other cases) are linked to a set of news articles from the period 2015-2017 that report on these cases. This dataset is augmented with perspective annotations from Minnema et al. (2022a). Concretely, there are a total of 400 sentences are available in gold annotations (averaged z-scored perception values from 240 participants), and 7,754 further sentences in silver annotations (annotated with the best-scoring model from Minnema et al. 2022a).

Using event metadata, we automatically extracted pairs of sentences $\langle L, H \rangle$, where L and H both reference the same GBV case, but respectively have a below-average (L) or above-average (H) level of perceived perpetrator blame on the perpetrator. Next, for a subset of 1,120 sentences from the combined gold-silver perspective dataset, we performed manual filtering to ensure that for each pair, L and H reference not only the same *case*, but also show substantial overlap in terms of the specific *events* within this case that they describe (e.g. the violence itself, the police investigation, conviction of a suspect, etc.). Since a sentence might be used multiple times, the alignment process yielded a set of 2,571 pairs (or 304 pairs if each sentence is used only once).

11.3 Methods

Due to the limited availability of task-specific parallel training data, we experiment with several existing text rewriting methods known to work in low-data settings: unsupervised, zero-shot, and few-shot methods, as shown in Figure 11.1.

11.3.1 Unsupervised Methods

We use the unsupervised method from Chapter 5 that does not require any task-specific parallel data, and two corresponding variants of this method that use metadata.

[2] Including cases of non-lethal violence, suspected femicide, and suicide.

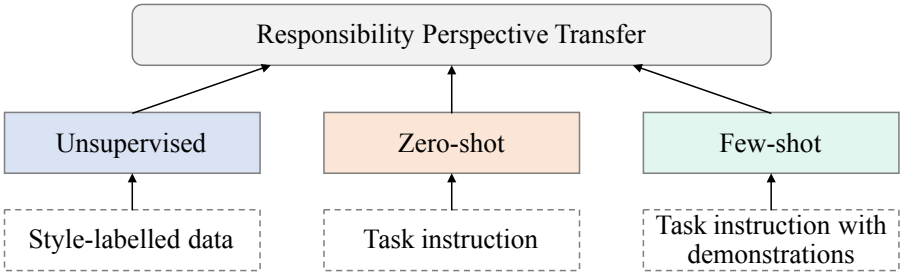

Figure 11.1: Overview of three different methods for responsibility perspective transfer.

Unsupervised mBART We train an unsupervised framework based on iterative back-translation (Hoang et al., 2018) method using non-parallel style-labelled data. The framework consists of two mBART-based models, one for each transfer direction, where outputs of one direction with source sentences are used to supervise the model in the opposite direction. All experiments are implemented atop Transformers (Wolf et al., 2020) using mBART-50 (Tang et al., 2021). We use the Adam optimiser with a polynomial learning rate decay, and a linear warmup of 100 steps for a maximum learning rate of 1e-4. We limit the maximum token length to 150 for both source and target sentences. To alleviate computational costs and catastrophic forgetting, we only update the parameters of the decoder, freezing the other parameters.

mBART + metadata A unique feature of our dataset is the availability of detailed metadata about the events. We made a selection of the properties likely to be most relevant for characterizing the event and assigning responsibility (names of the victim and perpetrator, type of victim-perpetrator relationship, murder weapon and location) and concatenated this metadata to the corresponding source sentence as input during training. We tried two order settings: ***source-meta*** and ***meta-source***. Preliminary experiments showed that concatenating only the event properties themselves, without including property names, produced the most promising results. For example: *"Trapani, Donna di 60 anni uccisa dall'ex marito — Anna Manuguerra, Antonino Madone, ex coniuge, arma da taglio, Nubio, casa"* ("Trapani: 60-year old woman killed by ex-husband — [victim name], [perpetrator name], ex-

Riscrivi la frase concentrandoti sul colpevole
[Source sentence]
– –
(EN: *Rewrite the sentence and concentrate on the culprit*\n [Source sentence])

Figure 11.2: Prompt template for naive zero-shot.

Riscrivi le seguenti frasi da low ad high.
Per high si intende che la colpa è attribuita interamente al killer.
Ecco alcuni esempi:
[Examples]
Riscrivi la seguente frase:
[Source sentence]
– –
(EN: *Rewrite the following sentences from low to high.*\n *'High' means that the blame is entirely put on the killer.*\n *Here are some examples:*\n[Examples]\n *Rewrite the following sentence:*\n [Source sentence])

Figure 11.3: Prompt template for naive few-shot.

spouse, cutting weapon, [town name], at home"). We use the same training setup as for the previous model.

11.3.2 Zero-/Few-shot Methods

We also experimented with using the *text-davinci-002* version of GPT-3 in a range of zero-shot and few-shot setups.

GPT-3: Naive implementation In our ***naive-zero*** setup, as shown in Figure 11.2, we use a simple prompt telling the model to rewrite the sentence with more focus on the perpetrator. Figure 11.3 shows the prompt template for our ***naive-few*** setup, which uses a similarly simple prompt along with a specific task definition and a set of ten low-high sentence pairs randomly sampled from the gold annotations.

GPT-3: Iterative few-shot A challenging factor for our naive few-shot ap-

iter-1
Ecco delle frasi che parlano di femminicidio.
Le frasi precedute dall'etichetta "Low:" tendono ad essere più brevi e non danno la colpa esplicita all'assassino, mentre le frasi precedute dall'etichetta "High:" tendono ad essere più dirette e a dare la colpa all'assassino.
[Examples]
[Source sentence]

---

iter-2
Ecco 10 frasi parlando di femminicidi. Le frasi cominciando con "Low:" tendono ad essere più brevi e non danno la colpa esplicita all'assassino, mentre le frasi cominciando con "High:" tendono ad essere più dirette e a dare la colpa all'assassino.
[Examples]
[Source sentence]

Figure 11.4: Prompt templates for iterative few-shot.

proach is that the 'natural' source-target pairs from our annotated data are not perfect minimal pairs, as they differ in perspective but also have some content differences. In an effort to use maximally informative pairs as few-shot examples, we designed an iterative process for compiling small curated sets of examples. Specifically, we designed an improved prompt method by giving a set of source-target pairs sampled from the gold annotations to the model and prompting it to explain the differences between the pairs. We discovered by accident that this yields a very plausible and concise task definition[3], and we reasoned that a definition generated by the model on the basis of real examples might be more informative as a prompt than a manually designed one. Therefore, we adapted the generated definition into a zero-shot prompt, and used that prompt model to generate target sentences

[3]The definition (slightly edited for grammar) is: *"Le frasi precedute dall'etichetta "Low:" tendono ad essere più brevi e non danno la colpa esplicita all'assassino, mentre le frasi precedute dall'etichetta "High:" tendono ad essere più dirette e a dare la colpa all'assassino."* ("The sentences preceded by "Low:" tend to be shorter and don't explicitly blame the murderer, while the sentences preceded by "High:" tend to be more direct and blame the murderer.")

for five source sentences sampled from the corpus. We then selected the best candidate from these to create a set of pairs with maximal perspective contrast and content overlap, to be used in a few-shot prompt. As shown in Figure 11.4, we kept both versions of the few-shot prompt, ***iter-1*** and ***iter-2*** in order to measure the combined effects of small difference in prompt, randomness in the generated candidates, and judgement differences in the selection of the best candidate.

11.4 Experiments

11.4.1 Evaluation Methods

The main goal of responsibility perspective transfer is to generate a sentence with the desired perspective that still has the same semantic content as the source sentence. We employ commonly used automatic metrics in text style transfer to assess the performance of different models on two aspects: style strength and content preservation (more details can be found in Chapter 3). In language generation tasks, human evaluation is always considered very valuable, although it is not always affordable. Here, we want to have at least a small-scale study with human subjects to assess the validation of the automatic metrics for our new task. Therefore, we also run a questionnaire-based evaluation with human participants.

Automatic Evaluation For assessing perspective quality (i.e. style strength), we used the best-performing perspective regressor from Minnema et al. (2022a) which is based on an Italian monolingual DistilBERT model (*BERTino*, Muffo and Bertino 2020). For content preservation, we use three popular text generation metrics: n-gram-based *BLEU* (Papineni et al., 2002) and *ROUGE* (Lin, 2004), as well as the neural-based model *COMET*[4] (Rei et al., 2020).

Human Evaluation In our human evaluation setup, participants were given an online survey with 50 blocks, each corresponding to one source sentence sampled from the dataset. In each block, participants rated: 1) the level of

[4]Although COMET is designed to also take input sentences into account, we do not use the input setting because COMET training input and output are different languages.

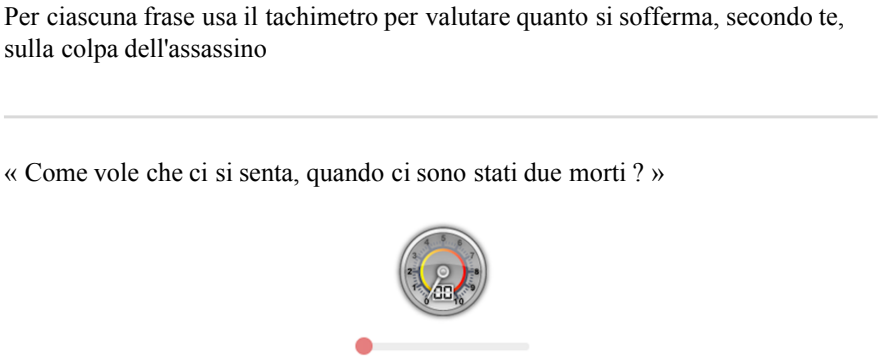

Figure 11.5: A screenshot of rating the perceived level of blame of the perpetrator.

perceived agent responsibility in each of the seven *target* candidates; 2) the level of *content preservation* of each target relative to the source. We also designed a separate, smaller questionnaire that asked the same questions about the few-shot examples used in ***iter-1*** and ***iter-2***. Figure 11.5 shows a screenshot of the annotation interface used to rate the perceived level of blame of the perpetrator.

The pool of invited participants was a group of people with mixed genders and backgrounds from the personal network of the authors. No remuneration was offered. Four invitees responded to the main questionnaire, and three invitees responded to the few-shot example questionnaire (all female, mean age: 46). The participants have different levels of education (from middle school to university) and live in different regions of Italy. Our evaluation study should be seen as a pilot, and larger-scale, more representative studies are planned for the future. The main aim of the pilot was to have a small-scale validation of our automatic metrics (taken from previous work and developed on the basis of a large-scale human study) and to test our evaluation setup (which questions to ask, etc.). The questionnaire was designed and distributed using Qualtrics.[5]

[5] https://www.qualtrics.com/

Perspective Model Dimension	R^2	Source	Target	mBART base	mBART src-m	mBART m-src	GPT-3 n-zero	GPT-3 n-few	GPT-3 iter-1	GPT-3 iter-2
"blames the murderer"	*0.61*	-0.511	0.445	-0.250	-0.188	**0.284**	-0.157	-0.375	**0.109**	-0.116
"caused by humans"	*0.60*	-0.228	0.362	-0.037	0.005	**0.371**	0.042	-0.095	**0.278**	0.076
"focuses on murderers"	*0.65*	-0.518	0.597	-0.184	-0.108	**0.567**	0.033	-0.349	**0.179**	-0.104
"focuses on victims"	*0.63*	0.019	-0.137	0.035	0.171	**0.173**	-0.123	0.178	**0.195**	0.097

Table 11.2: Automatic evaluation of perspective using the BERTino-based model from Minnema et al. (2022a). Scores are z-normalized (i.e. a score -1 or 1 means “one standard deviation below/above average”).

11.4.2 Automatic Results

Perspective Evaluation Following Minnema et al. (2022a), we distinguish between several perceptual dimensions using a perception regression model, as shown in Table 11.2. Our main dimension of interest is *blame on murderer*, but we also look at the two closely related dimensions of *cause* and *focus on murderer*. As shown by the R^2 scores, regression quality is decent for all of these dimensions. We observe that the source and target sentences have lower and higher blame scores respectively, which are also consistent on the two related dimensions, affirming that our testing data is of good quality in terms perspective aspect.

For all models, the perception scores of the predicted sentences are higher than those of the source sentences, with mBART/*meta-src* achieving the highest scores. This suggests that all models alter perceptions of responsibility to some extent. However, in virtually all cases, perception scores stay well below the target, and in many cases below the average level (zero). For mBART-based results, models with metadata perform better than the baseline, with *meta-src* reaching particularly high scores. Within the GPT-3 settings, zero-shot (*na-zero*), surprisingly, performs better than few-shot; (*na-few*), and *iter-1* yields the highest scores.

Content Preservation As shown in Table 11.3, when taking source sentences as the reference, three metrics show that the outputs have higher similarities to them than the target sentences. mBART/*base* has the highest scores, which (combined with the low perception scores of this model) suggests that the model tends to copy from the source sentence. Within the GPT-3 settings,

Metric	↔	Source	Target	mBART			GPT3			
				base	src-m	m-src	na-zero	na-few	iter-1	iter-2
BLEU	src	-	1.5	**72.5**	61.2	23.6	30.3	43.5	**48.9**	28.5
ROUGE	src	-	10.0	**80.8**	70.1	35.1	55.1	63.8	**65.9**	45.0
COMET	src	-	-121.6	**54.0**	25.7	-59.1	10.3	**53.8**	37.9	-5.8
BLEU	tgt	1.5	-	1.4	1.6	**2.4**	1.0	1.3	**1.4**	0.9
ROUGE	tgt	10.0	-	11.0	10.4	**13.2**	8.8	9.4	**9.8**	9.0
COMET	tgt	-117.5	-	-119.4	-117.8	**-100.2**	-109.0	**-104.5**	-105.7	-105.9

Table 11.3: Automatic content preservation metrics (BLEU, ROUGE, COMET), comparing generated sentences against source and gold target sentences.

		Perspective	Similarity	HM
mBART	base	2.14	**7.72**	3.34
	src-meta	2.50	6.78	3.65
	meta-src	**4.50**	3.62	**4.01**
GPT-3	na-zero	2.77	6.52	3.89
	na-few	2.08	**8.17**	3.31
	iter-1	3.57	7.97	**4.98**
	iter-2	**3.84**	6.60	4.85
Examples	for iter-1	**5.20**	**6.93**	**5.94**
	for iter-2	3.87	5.27	4.46

Table 11.4: Human evaluation results on model outputs and examples for few-shot. HM is the harmonic mean of perspective and similarity scores

iter-1 has the highest scores. Using instead the target sentences as reference, we see that all scores are very close, with mBART/*meta-src*) reaching the best performance, followed by GPT-3/*na-few* and GPT-3/*iter-1*.

11.4.3 Human-based Results

Inter-annotator agreement In our experiments, we found reasonably high levels of inter-annotator agreement (Spearman's rank correlation between pairs of annotators). Correlations ranged between 0.3-0.6 (blame) and 0.4-0.6 (similarity) with high levels of significance ($p < 0.0001$).

Results Table 11.4 reports the results of our human evaluation study. We see

that mBART/*meta-src* is the best overall model on perspective, but has poor performance on the similarity aspect. Meanwhile, GPT3/*na-few* achieves the highest score on similarity but the lowest score in terms of perspective, and its overall performance is lower than that of GPT3/*na-zero*. GPT3/*iter-1* has the best overall performance with an HM of 4.98.

11.5 Analysis and Discussions

11.5.1 Case Styudy

Table 11.5 shows two sets of example outputs generated by mBART and GPT-3.[6] While hand-picked, these examples show that both models are capable of generating sentences that increase responsibility while trying to preserve content. However, they also highlight a key challenge: what if the source sentence lacks details about the event? The mBART model has access to event metadata and uses this effectively in Example 1 to produce a sentence that stays close to the source but with details from the metadata filled in (though with rather clunky sentence structure). In Example 2, instead, it produces a sentence that is factually correct but also loses most of the information from the source sentence. On the other hand, GPT-3, without access to metadata, often 'invents' information. This is evident in the second example, in which it faithfully preserves the source sentence and increases the level of blame by adding plausible but (partially) incorrect information about the crime.

11.5.2 Limitations

We introduced the new task of responsibility perspective transfer and provided initial data collection and modelling for a specific domain (news about gender-based violence) and language (Italian). The main limitation is the (mBART) models that we trained and the prompts (GPT-3) that we designed are specific to this domain and language and cannot be applied 'out-of-the-box' in other contexts. However, all of our modelling setups require no or limited parallel data and make use of readily available existing models, so we believe the general approach to be easily transferrable to other domains.

[6]We include generations from the overall best-performing model from each category.

Models	Sentences
Source	*Provaglio d'Iseo, donna trovata morta in casa : si sospetta il compagno* "[town name], woman found dead at home: [her] partner is suspected"
Target	*Donna massacrata in casa nel Bresciano, il convivente fermato per omicidio volontario* "Woman massacred at home in [area name]: [her] cohabitant [partner] arrested for voluntary homocide"
mBART (meta-src)	*Un uomo, Elio Cadei, è stato fermato per l'omicidio di Simona Simonini, la donna trovata morta in casa: si sospetta il compagno di lei.* "A man, Elio Cadei, has been arrested for the murder of Simona Simonini, the woman found dead at home: the partner of her is suspected"
GPT-3 (iter-1)	*Provaglio d'Iseo, donna trovata morta in casa: il compagno è il principale sospettato* "[town name], woman found dead at home: [her] partner is the principal suspect"
Source	*«Come vuole che ci si senta, quando ci sono stati due morti?»* "«How do you want us to feel, when there have been two deaths?»"
Target	*Albenga, uccide l'ex compagna a coltellate e poi si toglie la vita Loredana e l'incubo-stalker foto* "[town name], [someone] kills [his] ex-partner by stabbing and then takes his life[.] Loredana and the nightmare stalker[, see] foto"
mBART (meta-src)	*Un uomo, Mohamed Aziz El Mountassir, di 43 anni, ha ucciso la moglie, Loredana Colucci, di 44, con una coltellata al collo.* "A man, Mohamed Aziz El Mountassir, 43 years old, has killed his wife, Loredana Colucci, 44 years old, with a knife to [her] neck."
GPT-3 (iter-1)	*«Come vuole che ci si senta, quando ci sono stati due morti causati da un uomo che ha ucciso prima la sua ex moglie e poi la sua nuova compagna?* "How do you want us to feel, when there have been two deaths caused by a man who has first killed his ex-wife and then his new partner?"

Table 11.5: Example source sentences and their corresponding target sentences and model outputs.

Another limitation comes from the fact that we used GPT-3: the model is closed-source and can only be accessed with a paid subscription to the OpenAI API (`https://beta.openai.com/`). This has consequences for reproducibility for several reasons. First of all, we do not have access to the exact technical specifications of the model or to the training data that was used. The GPT-3 models are regularly updated (at the time of our experiments, *text-davinci-002* was the most recent available version), but limited information is available about what distinguishes each version from the previous ones or from the original model introduced in Brown et al. (2020). Moreover, access to the API is controlled by OpenAI and could be closed at any time at the company's discretion; the API is currently quite accessible with no waiting list and a reasonably generous free trial, but the rates (paid in USD) might not be affordable for researchers outside of institutions in high-income countries, and not all researchers might be comfortable agreeing to the company's terms and conditions. Finally, the generative process involves a degree of randomness, and through the API it is not possible to fixate the model's random seed, meaning that the model produces different predictions every time it is called, even when using exactly the same prompt.

11.5.3 Ethics Statement

We see three important ethical considerations for the content of this chapter. The first consideration is related to the use of large language models. Apart from the reproducibility limitations resulting from the use of GPT-3, there are more general questions surrounding the use of GPT-3 and similar models, for example the high energy usage and resulting carbon emissions, and societal questions around the oligopoly on state-of-the-art language models that is currently in the hands of a handful of large companies.

The second consideration relates to the task and dataset that we introduce: the data could unfortunately be used maliciously, for example, to build a system that puts more focus on the victim. While we cannot fully prevent such uses once our data are made public, we do hope that writing about risks explicitly and also raising awareness of this possibility in the general public are ways to contain the effects of potential harmful uses. We are open to any

discussion and suggestions to minimise such risks. On the other hand, while we see perspective transfer models as a valuable tool for studying how language 'frames' (social) reality that could also have practical applications, for example in journalism, we strongly believe that any such applications must be approached with extreme care. The models that we introduce are scientific analysis tools that could be used to suggest alternative viewpoints on an event, but we believe that generations should *not* be seen as necessarily reflecting a 'true' or 'better' perspective, and should not used in a prescriptive way (i.e. used to tell someone how to write). We believe that the authors (journalists or others) of any text ultimately bear exclusive responsibility for the views, perspectives and (implicit) values expressed in it, and should be careful in making use of texts (re-)written by computers, such as the ones produced by our proposed models.

Finally, we are aware that our task domain (femicide/gender-based violence) is a societally and emotionally loaded topic, and that the texts contained in our dataset and produced by our models might be disturbing. In particular, in some cases, models may produce graphic descriptions of violence and/or produce questionable moral judgements (e.g. we have occasionally seen statements such as "the perpetrator of this horrible crime does not have the right to live" spontaneously produced by some of the models), and potential users of applications of the model should be aware of this. For the purposes of this chapter, the only people external to the research team who have been extensively exposed to model outputs were the annotators in our human evaluation study. In the introduction page of our online questionnaire, annotators were warned about the sensitive nature of the topic and advised that they could stop their participation at any time if they felt uncomfortable and could contact the authors with any questions. Prior to running the online questionnaire we have requested and obtained ethical approval by the Ethical Review Committee of our research institution.

11.6 Conclusion

In this chapter we proposed responsibility perspective transfer as a new task and introduced a dataset and models for applying this task to Italian news

reporting about femicides. Our dataset contains a limited amount of quasi-aligned pairs that proved useful for evaluation and few-shot learning. We experimented with two modeling approaches: unsupervised mBART (with or without enriching the input with metadata) and zero-shot/few-shot learning with GPT-3.

Our human and automatic evaluations suggest *GPT-3/iter-1* as the best overall model, with a relatively high level of responsibility placed on the perpetrator and a good degree of content preservation. For the latter, most models score at least 6/10 on average on the human survey. The perspective change itself has also been achieved by our models, with substantially increased levels of perceived perpetrator blame compared to the source, but there is still much room for improvement: none of the models comes close to having the same level of blame as the target sentences do, and in the human evaluation survey no model achieves a 'blame score' of more than 4.5/10. The main obstacle for future improvements seems to lie with the lack of truly parallel data; however, our GPT-3-based iterative approach of creating minimal pairs seems to have worked quite well, and might be further exploited on a larger scale.

In the next chapter (Chapter 12), we will consider a special rewriting task: meaning-to-text generation (i.e. producing a text from a meaning representation), and explore multilingual language-meaning modelling for this task.